Trabajo Práctico Final IECD

Test de Wilcoxon de Rango Signado

Introducción a la Estadística y Ciencia de Datos

Diciembre 2024 - 2C

Autor	Correo electrónico
Falczuk, Noelia	noefalczuk@gmail.com
Moroni, Giancarlo	moronigiancarlo1@gmail.com
Schmidt, Tomás	tomas.schmidt2004@gmail.com

Índice

Motivación

El presente trabajo práctico tiene como propósito principal profundizar los conocimientos en la programación orientada a objetos (OOP) en R, principalmente bajo el sistema S3. A través de este trabajo, se busca comprender en detalle la estructura y funcionamiento de las clases y métodos S3, así como la lógica subyacente de comandos esenciales, tales como s3 dispatch y class.

Además, este trabajo está orientado a introducir y analizar en profundidad el test de Wilcoxon, un método estadístico no paramétrico. Los objetivos asociados incluyen:

Comprender el pivote que utiliza el test para su cálculo. Analizar la distribución del estadístico bajo la hipótesis nula (H0) y los métodos de aproximación disponibles, cuando las muestras son grandes. Entender como funciona el test en diseños de muestras únicas y diseños apareados. Explorar la distribución del estadístico bajo la hipótesis alternativa (H1) vía boostrap.

Resumen

El objetivo principal de este trabajo práctico es estudiar el *test de rango signado de Wilcoxon*, un test no paramétrico utilizado para la mediana de una distribución simétrica. Se pretende implementar este test de manera compatible con la función nativa `wilcox.test` de R y evaluar su potencia frente a alternativas puntuales utilizando el método de bootstrap.

En primer lugar, se introduce el concepto de Programación Orientada a Objetos (OOP) en R, en S3. Se aprenderá como funcionan las clases de objetos en R.

Luego, se analiza el test de Wilcoxon y su distribución bajo H0. Se implementa el estadístico de prueba basado en rangos signados, que es esencial para este test. Además, se realiza un análisis de la distribución exacta del estadístico T^+ bajo la hipótesis nula y se estimará la potencia del test bajo la hipótesis alternativa usando el método de bootstrap paramétrico.

Finalmente, se compara el test de Wilcoxon con otros tests, como el test t, el teat para normales visto durante el curao y el test del signo, evaluando sus respectivas potencias y analizando cuál de ellos resulta ser más potente para nuestras hipótesis.

1. OOP en R

1.1. Pregunta 1

En R hay cinco clases básicas de vectores: character (letras), numeric (números reales), integer (números enteros), complex (números complejos) y logical (verdadero/falso o TRUE/FALSE). Los vectores en R solo pueden contener elementos de una misma clase. Por lo tanto, cuando un vector contiene elementos de diferentes clases, R realiza automáticamente una conversión para evitar errores, convirtiendo los elementos a una misma clase. Este proceso se denomina coerción.

Sin embargo, en algunos casos, la coerción no es posible para todos los elementos y R introduce NA (Not Available) para representar datos faltantes.

- Caso 1: `c(T, F)` Este vector contiene solo valores lógicos (`TRUE` y `FALSE`), por lo que su clase es `logical`, pues todos los elementos son de la clase logical entonces el vector también será de esta clase.

- Caso 2: `c(T, F, 1)` Aquí se mezclan valores lógicos (`TRUE` y `FALSE`) con un número (`1`). Debido a la coerción, R convierte todos los elementos a la clase `numeric` pues es la clase más general que puede contener a ambos.
- Caso 3: `c(T, F, 1, "1")` En este caso, los valores incluyen un lógico (`TRUE`), un número (`1`) y una cadena de caracteres (`"1"`). Como la clase `character` es la más general, R convierte todos los elementos a esta clase.

1.2. Pregunta 2

La función density pertenece a la clase function, ya que es una herramienta utilizada para calcular la estimación de la densidad. Por otro lado, cuando se ejecuta density(1:500), el resultado pertenece a la clase density.

La diferencia radica en que, en el primer caso, se está haciendo referencia directamente a la función density(), la cual es un objeto de clase function que, al ser aplicada, computa la estimación de la densidad. En cambio, al llamar a density(1:500), se aplica la función density a los datos del vector 1:500, generando un objeto de clase density que contiene la estimación de la densidad para esos datos.

1.3. Pregunta 3

En R, print() es una función genérica en S3. Es decir, cómo funciona print cambia dependiendo de la clase del objeto que se le pase. Esto se debe a que S3 es un sistema de programación orientado a objetos que permite definir comportamientos específicos para los objetos que se crean. Al ejecutar $methods('print')$ se podrán observar qué clases sabe despachar el genérico print. Sin embargo, la cantidad y quiénes son esas clases depende de las liberías cargadas y de la versión de R que se tenga; paquetes adicionales o objetos creados permitirían despachar nuevas clases. En nuestro caso despacha 266 clases.

En cuanto a density, al ejecutar $methods('density')$ se podrá observar la lista de métodos con los que cuenta density. En nuestro caso, density cuenta con otros cinco (5) métodos más además de $plot : coerce, initialize, print, show$ y $slotsFromS3$.

1.4. Pregunta 4

Nuevamente, ocurre algo similar a lo que sucedía en la pregunta dos (2). test_t es un objeto de clase htest, por lo tanto al imprimir test_t en pantalla observamos una salida donde los datos se imprimen acompañados de frases que explican qué son. El print sabe despachar a la clase htest. Al hacer $unclass(test\_t)$, este comando devuelve una copia de su argumento con la clase atributo removida. Es decir, devuelve simplemente los datos que se obtienen con la función test_t pero no en el formato htest sino como lista. Por esto, al ejecutar $class(unclass(test\_t))$ obtenemos que la clase es *lista*.

2. Test de Wilcoxon de rango signado para una muestra

2.1. Diseños experimentales apareados

2.1.1. Pregunta 5

Para probar que la distribución de $D = X - Y$ es simétrica alrededor del cero es suficiente con probar que

$$P(T - C \leq x) = P(T - C \geq -x) \iff P(T - C \leq x) = P(C - T \leq x)$$

Sea h una función boreliana, vale lo siguiente:

$$P(T - C \leq x) = P(h(C,T) \leq x) \underset{(1)}{=} P(h(W_1, W_2) \leq x) \underset{(2)}{=} P(h(T,C) \leq x) = P(C - T \leq x)$$

$$(1)\ \forall (W_1, W_2)\colon (C,T) \sim (W_1, W_2) \implies P(h(C,T) \leq x) = P(h(W_1, W_2) \leq x)$$

Para (2) utilizamos (1) con $(W_1, W_2) = (T, C)$.

Por lo tanto D es simetrica alrededor del cero (0).

2.2. Test de Wilcoxon

2.2.1. Pregunta 6

Sea

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{si } x \in (0, \infty), \\ 0 & \text{si } x \notin (0, \infty). \end{cases}$$

$$\mathbb{1}_B(x) = \begin{cases} 1 & \text{si } x \in (-\infty, 0), \\ 0 & \text{si } x \notin (-\infty, 0). \end{cases}$$

$$\mathbb{1}_C(x) = \begin{cases} 1 & \text{si } x \neq 0, \\ 0 & \text{si } x = 0. \end{cases}$$

$$T^+ + T^- = \sum_{i=1}^{n} \mathbb{1}_A(x_i) \cdot R_i + \sum_{i=1}^{n} \mathbb{1}_B(x_i) \cdot R_i = \sum_{i=1}^{n} \mathbb{1}_C(x_i) \cdot R_i = \frac{n(n+1)}{2}$$

$$T^+ - T^- = \sum_{i=1}^{n} \mathbb{1}_A(x_i) \cdot R_i - \sum_{i=1}^{n} \mathbb{1}_B(x_i) \cdot R_i = \sum_{i=1}^{n} \mathbb{1}_A(x_i) R_i + \sum_{i=1}^{n} (-1) \cdot \mathbb{1}_B(x_i) \cdot R_i =$$

$$= \sum_{i=1}^{n} (\mathbb{1}_A(x_i) - 1 \cdot \mathbb{1}_B(x_i)) R_i = \sum_{i=1}^{n} Signo(x_i) \cdot R_i = T$$

De esta manera:

$$T^+ = \frac{n(n+1)}{2} - T^- = \frac{T + \frac{n(n+1)}{2}}{2}$$

$$T^- = \frac{n(n+1)}{2} - T^+ = \frac{\frac{n(n+1)}{2} - T}{2}$$

$$T = T^+ - T^- = 2 \cdot T^+ - \frac{n(n+1)}{2} = \frac{n(n+1)}{2} - 2 \cdot T^-$$

2.3. Distribución de T+ bajo la hipótesis nula

2.3.1. Pregunta 7

Para mostrar que $|X_i|$ y que $s(X_i)$ son independientes es suficiente con mostrar que la función de probabilidad se factoriza. Es decir, es suficiente con mostrar lo siguiente:

- $P(s(X_i) = 1, |X_i| \leq x) = P(s(X_i) = 1)P(|X_i| \leq x)$
- $P(s(X_i) = -1, |X_i| \leq x) = P(s(X_i) = -1)P(|X_i| \leq x)$
- $P(s(X_i) = 0, |X_i| \leq x) = P(s(X_i) = 0)P(|X_i| \leq x)$

Veamos el primer caso. El segundo caso es análogo y el tercero es trivial pues $P(s(X_i) = 0) = 0$

$$\begin{aligned} P(s(X_i) = 1,\, |X_i| \leq x) &= P(X_i > 0, -x \leq X_i \leq x) = P(0 < X_i \leq x) \\ &= \frac{1}{2} \cdot 2 \cdot P(0 < X_i \leq x) \underset{(1)}{=} \frac{1}{2} \cdot P(|X_i| \leq x) \\ &= P(s(X_i) = 1)P(|X_i| \leq x) \end{aligned}$$

$$\begin{aligned} P(s(X_i) = -1,\, |X_i| \leq x) &= P(X_i < 0, -x \leq X_i \leq x) = P(x \geq X_i < 0) \\ &= \frac{1}{2} \cdot 2 \cdot P(x \geq X_i < 0) \underset{(1)}{=} \frac{1}{2} \cdot P(|X_i| \leq x) \\ &= P(s(X_i) = -1)P(|X_i| \leq x). \end{aligned}$$

Nota: La igualdad marcada como (1) se debe a que, bajo H_0, las X_i son simétricas alrededor de 0.

2.3.2. Pregunta 8

Como $(R_1, \ldots, R_n)$ es función de $(|X_1|, \ldots, |X_n|)$, y para todo par $(s(X_i), |X_i|)$ y $(s(X_j), |X_j|)$, $j, i = 1, \ldots, n$ con $i \neq j$ vale que son independientes, es suficiente mostrar que $s(X_i)$ y $|X_i|$ son independientes. La independencia de $s(X_i)$ y $|X_i|$ ya fue demostrada en el inciso anterior. Así, como los signos son independientes de los módulos, son independientes de los rangos pues son función de los módulos. También, son independientes de los antirangos pues estos son función de los rangos que son función de los módulos. Finalmente, $(R_1, \ldots, R_n)$ y $(D_1, \ldots, D_n)$ son independientes de $(s(X_1), \ldots, s(X_n))$ por ser ambos función de $(|X_1|, \ldots, |X_n|)$.

Aclaración: $(s(X_i), |X_i|)$ y $(s(X_j), |X_j|)$ son independientes pues las muestras X_i son tomadas de manera independiente.

2.3.3. Pregunta 9

Para probar que bajo $H_0 : \theta = 0, F \in \Omega_s$, las v.a. $W_j = \mathbb{1}\{X_{D_j} > 0\}$ distribuyen según

$$W_1, \ldots, W_n \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$$

Se observa que:

$$W_j = \begin{cases} 1 & \text{si } X_{d_j} > 0 \\ 0 & \text{sino .} \end{cases}$$

y se ve que $P(W_i = 1) = P(X_{d_i} > 0) = P(W_i = 0) = P(X_{d_i} < 0) = \frac{1}{2}$. De esta manera, se deduce que $W_i \sim Be(\frac{1}{2})$

Para probar la independencia, es suficiente con probar que $P(W = w) = \prod_{j\in[n]} P(W_j = w_j)$ donde $w \in \{0,1\}^n$.

$$P(W_1 = w_1, \ldots, W_n = w_n) \underset{(0)}{=} \sum_d P\left(\mathbb{1}\{X_{D_1} > 0\} = w_1, \ldots, \mathbb{1}\{X_{D_n} > 0\} = w_n \mid D = d\right) P(D = d)$$

$$\underset{(1)}{=} \sum_d P\left(\mathbb{1}\{X_{d_1} > 0\} = w_1, \ldots, \mathbb{1}\{X_{d_n} > 0\} = w_n \mid D = d\right) P(D = d)$$

$$\underset{(2)}{=} \sum_d P\left(\mathbb{1}\{X_{d_1} > 0\} = w_1, \ldots, \mathbb{1}\{X_{d_n} > 0\} = w_n\right) P(D = d)$$

$$\underset{(3)}{=} \sum_d P\left(\mathbb{1}\{X_{d_1} > 0\} = w_1\right) \ldots P\left(\mathbb{1}\{X_{d_n} > 0\} = w_n\right) P(D = d)$$

$$= \left(\frac{1}{2}\right)^n \sum_d P(D = d) = \left(\frac{1}{2}\right)^n$$

Por lo tanto, $P(W_1 = w_1, \ldots, W_n = w_n) = \prod_{i=1}^{n} P(W_i = w_i)$ y $\quad P(W_i = w_i) = \frac{1}{2}$.

(0) Realizamos proba total.
(1) Por definición de probabilidad condicional
(2) Pues $\mathbb{1}\{X_{d_n} > 0\}$ depende únicamente del signo de X_{d_n}
(3) Pues los signos de X_i y X_j son independientes cuando $i \neq j$
D = { Conjunto de todos los posibles valores que puede tomar el rango }

2.4. Distribución exacta de T+

2.4.1. Pregunta 10

Para reproducir la tabla se uso R. El código se encuentra en el archivo 'codigo-Falczuk-Moroni-Schmidt.R'. La tabla obtenida para n = 5 es la que se muestra a continuación, junto a la original para n = 4.

2.4.2. Pregunta 11

Para probar que T^+ es simétrica alrededor de $\frac{n(n-1)}{4}$, es suficiente con mostrar que

$$P(T^+ = \frac{n(n+1)}{2} - K) = P(T^+ = K) \quad \forall\, K\colon 0 \leq K \leq \frac{n(n+1)}{2}.$$

Imágenes del conjunto de datos

t	$S_{4,t}$	$\#S_{4,t}$	$p_4(t)$
0	$\{\emptyset\}$	1	1/16
1	$\{\{1\}\}$	1	1/16
2	$\{\{2\}\}$	1	1/16
3	$\{\{3\},\{1,2\}\}$	2	2/16
4	$\{\{4\},\{1,3\}\}$	2	2/16
5	$\{\{1,4\},\{2,3\}\}$	2	2/16
6	$\{\{2,4\},\{1,2,3\}\}$	2	2/16
7	$\{\{3,4\},\{1,2,4\}\}$	2	2/16
8	$\{\{1,3,4\}\}$	1	1/16
9	$\{\{2,3,4\}\}$	1	1/16
10	$\{\{1,2,3,4\}\}$	1	1/16

Figura 1: n = 4

	T	S	Cantidad	Probabilidad
1	0	{ }	1	0.03125
2	1	{1}	1	0.03125
3	2	{2}	1	0.03125
4	3	{3}{1, 2}	2	0.06250
5	4	{4}{1, 3}	2	0.06250
6	5	{5}{1, 4}{2, 3}	3	0.09375
7	6	{1, 5}{2, 4}{1, 2, 3}	3	0.09375
8	7	{2, 5}{3, 4}{1, 2, 4}	3	0.09375
9	8	{3, 5}{1, 2, 5}{1, 3, 4}	3	0.09375
10	9	{4, 5}{1, 3, 5}{2, 3, 4}	3	0.09375
11	10	{1, 4, 5}{2, 3, 5}{1, 2, 3, 4}	3	0.09375
12	11	{2, 4, 5}{1, 2, 3, 5}	2	0.06250
13	12	{3, 4, 5}{1, 2, 4, 5}	2	0.06250
14	13	{1, 3, 4, 5}	1	0.03125
15	14	{2, 3, 4, 5}	1	0.03125
16	15	{1, 2, 3, 4, 5}	1	0.03125

Figura 2: n = 5

Sea $\omega_j^{(a)} \in \{0,1\}^n$ tal que $\sum_{j\in[1:n]} \omega_j^{(a)} \cdot j = K$ con $0 < K < \frac{n(n+1)}{2}$.

Y sea $\omega_j^{(b)} \in \{0,1\}^n$ tal que $\sum_{j\in[1:n]} \omega_j^{(b)} \cdot j = \frac{n(n+1)}{2} - K$ con $0 < K < \frac{n(n+1)}{2}$.

(1) Veamos que $\omega_j^{(a)} = \omega_j^{(c)} - \omega_j^{(b)}$ donde $\omega_j^{(c)} \in \{1\}^n$ y cumple $\sum_{j\in[1:n]} \omega_j^{(c)} \cdot j = \frac{n(n+1)}{2}$

$$\omega_j^{(a)} = \omega_j^{(c)} - \omega_j^{(b)} \Leftrightarrow \omega_j^{(a)} + \omega_j^{(b)} = \omega_j^{(c)} \Leftrightarrow \sum_{j\in[1:n]} \omega_j^{(a)} \cdot j + \sum_{j\in[1:n]} \omega_j^{(b)} \cdot j = \frac{n(n+1)}{2}$$

Ahora, veamos que podemos reescribir a T_a^+ como:

$$T_a^+ = \sum_{j\in[1:n]} \omega_j^{(a)} \cdot j \underset{(4)}{=} \frac{n(n+1)}{2} - \sum_{j\in[1:n]} (1 - \omega_j^{(a)}) \cdot j \underset{(1)}{=} \frac{n(n+1)}{2} - \sum_{j\in[1:n]} \omega_j^{(b)} \cdot j$$

Y a T_b^+ como:

$$T_b^+ = \sum_{j\in[1:n]} \omega_j^{(b)} \cdot j \underset{(4)}{=} \frac{n(n+1)}{2} - \sum_{j\in[1:n]} (1 - \omega_j^{(b)}) \cdot j \underset{(1)}{=} \frac{n(n+1)}{2} - \sum_{j\in[1:n]} \omega_j^{(a)} \cdot j$$

(4) Al total se le resta los rangos que no estas sumando.

(3) Ademas, T_a^+ y T_b^+ son identicamente distribuidas pues existe una función h : $\mathrm{h}(\omega) = \sum_{j\in[1:n]} \omega_j \cdot j$, tal que $T_a^+ = \mathrm{h}(\omega^a)$ y $T_b^+ = \mathrm{h}(\omega^b)$ y $\omega^a \sim \omega^b$ (Teorema 1, pregunta 5)

Por otro lado,
Sea $K \in 0 \leq K \leq \frac{n(n+1)}{2}$ quiero ver que $P(T_a^+ = \frac{n(n+1)}{2} - K) = P(T_b^+ = K)$:

$$P(T_a^+ = \frac{n(n+1)}{2} - K) = P(T_b^+ = K)$$

$$\Leftrightarrow P(\frac{n(n+1)}{2} - \sum_{j\in[1:n]} \omega_j^b \cdot j = \frac{n(n+1)}{2} - K) = P(T_b^+ = K)$$

$$\Leftrightarrow P(\sum_{j\in[1:n]} \omega_j^b \cdot j = K) \underset{(2)}{=} P(T_b^+ = K)$$

(2) Vale por definición de T_b^+

Uniendo ambas cosas:

$$P(T_a^+ = \frac{n(n+1)}{2} - K) = P(T_b^+ = K) \underset{(3)}{=} P(T_a^+ = K)$$

Finalmente,

$$P(T_a^+ = \frac{n(n+1)}{2} - K) = P(T_a^+ = K) \ \forall \ K\colon 0 \leq K \leq \frac{n(n+1)}{2}.$$

2.4.3. Pregunta 12

La función que implementa la recursión $u_n(t)$ se encuentra en el archivo 'codigo-Falczuk-Moroni-Schmidt.R'

2.4.4. Pregunta 13

Las funciones que implementan la dTmas() y pTmas() se encuentran en el archivo 'codigo-Falczuk-Moroni-Schmidt.R'

2.4.5. Pregunta 14

Las funciones que implementan el test mi.wilcox.test() se encuentra en el archivo 'codigo-Falczuk-Moroni-Schmidt.R'

2.5. Distribución asintótica del test

2.5.1. Pregunta 15

Bajo H0, se calcula $E(T^+)$ y $Var(T^+)$ de la siguiente manera:

Dado que:

$$T^+ = \sum_{j=1}^{n} jW_j, \quad \text{bajo } H_0,$$

la $E(T^+)$ se calcula como:

$$E(T^+) = \sum_{j=1}^{n} j \cdot E(W_j) \underset{(1)}{=} \sum_{j=1}^{n} j \cdot \frac{1}{2} = \frac{n(n+1)}{4}.$$

La $V(T^+)$ se calcula como:

$$V(T^+) = \sum_{j=1}^{n} j^2 \cdot V(W_j) \underset{(1)}{=} \sum_{j=1}^{n} j^2 \cdot \frac{1}{4} = \frac{n(n+1)(2n+1)}{24}.$$

(1) $W_j \sim \text{Be}\left(\frac{1}{2}\right)$ e independientes (i.i.d), de la pregunta 9.

2.5.2. Pregunta 16

Para encontrar la distribución asintótica de T^+ no es posible usar la versión del Teorema Central del Límite que conocemos pues, a pesar de que es una combinación lineal de variables independientes e idénticamente distribuidas entre sí (las W_j), cada una está pesada por un coeficiente distinto (los $j \in [1:n]$). Por esto, usamos el TCL de Lindeberg.

En nuestro caso,

$$T^+ = \sum_{j=1}^{n} jW_j$$

entonces, tomando $V_i = W_i - \frac{1}{2}$ y $a_i = i$ y verificando las condiciones del teorema, se obtiene que sea $S = \sum_{i=1}^{n} \frac{a_i V_i}{\sqrt{n}}$, entonces:

$$\frac{S}{\sqrt{Var(S)}} \xrightarrow{d} Z \sim N(0,1).$$

Verifiquemos entonces las condiciones del teorema.

- $E(V_j) = 0$

$$E(V_j) = E(W_j - 1/2) = 0$$

- $Var(W_j) = \sigma^2$

$$Var(V_j) = Var(W_j - 1/2) = Var(W_j) = 1/4$$

- $\frac{max_i(|a_i|)}{\sqrt{\sum_{j=1}^{n} j \cdot a_i^2}} \xrightarrow{n\to\infty} 0$

$$\frac{max_i(|a_i|)}{\sqrt{\sum_{j=1}^{n} j \cdot a_i^2}} = \frac{n}{\sqrt{\frac{n\cdot(n+1)\cdot(2n+1)}{6}}} < \frac{\sqrt{6}\cdot n}{\sqrt{2n^3}} = \sqrt{\frac{3}{n}} \xrightarrow{n\to\infty} 0$$

Finalmente,

$$\frac{S}{\sqrt{Var(S)}} = \frac{\frac{\sum_{j=1}^{n} jV_j}{\sqrt{n}}}{\sqrt{Var(\frac{\sum_{j=1}^{n} jV_j}{\sqrt{n}})}} = \frac{\sum_{j=1}^{n} j(W_j - \frac{1}{2})}{\sqrt{n \cdot Var(\frac{\sum_{j=1}^{n} jV_j}{\sqrt{n}})}} =$$

$$\frac{\sum_{j=1}^{n} jW_j - \sum_{j=1}^{n} j \cdot \frac{1}{2}}{\sqrt{Var(\sum_{j=1}^{n} jV_j)}} \underset{(1)}{=} \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{d} Z \sim N(0,1).$$

$$(1)\ Var(\sum_{j=1}^{n} j \cdot V_j) = Var(\sum_{j=1}^{n} j \cdot (W_j - 1/2)) =$$

$$\sum_{j=1}^{n} j^2 \cdot Var(W_j - 1/2)) = \frac{1}{4} \cdot \sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{24}$$

O sea que,

$$T^+ \xrightarrow{d} Z \sim N(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}).$$

2.5.3. Pregunta 17

La función que permite graficar la distribución exacta de u^+ se encuentra en el archivo 'codigo-Falczuk-Moroni-Schmidt.R'

Los gráficos obtenidos son los que se muestran a continuación:

Figura 3

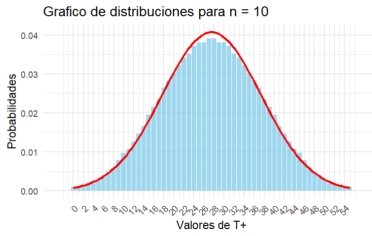

Figura 4

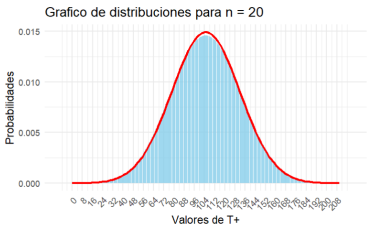

Figura 5

Se observa que los gráficos dan una buena intuición de cómo debería ser la distribución de T+. Es decir, todos los gráficos siguen la forma de la distrubución, a grandes rasgos, pero la continuidad la vamos obteniendo a medida que n crece. Cuando n = 4 la distribución exacta difiere de la calculada especialmente en las colas y en el centro. Cuando n es 10, podemos ver que mejora la aproximación en las colas pero todavía difiere en el centro y en las zonas medias entre el centro y las colas. Finalmente, para n = 20, la aproximación a la distribución exacta es muy buena ya que parecería coincidir razonablemente. Es decir, la distribución asintótica que aproxima a la distribución exacta parecería ser razonable para n >= 20.

Lo que esta sucediendo es que estamos tratando de aproximar la distribución de una función discreta por una función continua, lo que hace que las probabilidaes no sean exactamente comparables. Por esto, se peude realizar una corrección por continuidad.

La corrección por continuidad se basa en lo siguiente:

$$P[X_{discreta} = x] = P[x - 0{,}5 \leq X_{continua} \leq x + 0{,}5]$$

$$\textit{siempre y cuando } X_{discreta} \xrightarrow{d} X_{continua}.$$

2.6. Distribución bajo la alternativa vía bootstrap

2.6.1. Pregunta 18

Para llevar a cabo la resolución de este ejercicio fue necesario primero calcular el valor de K* que maximice la potencia del test $\phi_w(x) = \mathbb{1}_{\{T+ \geq k*\}}$. Por lo tanto, se calculó cuál es el cauntil 0.05 de la distribución de T+. Aprovechando que la distribución es discreta, vamos a ir calculando la P(T+ = t) desde el final al principio y sumando cada valor de probabilidad. Al encontrar el t tal que la suma de las probas supere 0,05, NO lo vamos a considerar y nos vamos a quedar con el t anterior a él. Luego este será el valor de k* que maximice la potencia (manteniendo el nivel del test) por el 'trade off' entre la probabilidad de error de tipo 1(Error I) y la probabilidad de error de tipo 2(Error II), ya que mientras más grande sea el valor de la P(Error I), más chica será la P(Error II). Por lo tanto, cómo la potencia del test bajo H1 es 1 - P(Error II), mientras más grande sea la P(Error I), mayor será la potencia del Test bajo H1.

Una vez encontrado k*, estimamos por bootstrap paramétrico la potencia del test. La misma nos dio un valor de 0.9271 lo que implica que P(Error II) = 1-0.9271 = 0.0729.

2.6.2. Pregunta 19

DISCLAIMER: El ejercicio fue pensado antes de la consigna de que σ era conocido.

Necesitamos poder calcular la potencia para el siguiente test, pensado con σ desconocido.

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu > 0.$$

$$T = \frac{\sqrt{n}\,\overline{X}_n}{S}$$

$$\phi_t(x) = \mathbb{1}_{\{T \geq k_\alpha\}}$$

$$k_\alpha = t_{\alpha,n-1} \text{ donde } P_{\theta_0}(T \geq t_{\alpha,n-1}) = 0{,}05$$

Veamos como se distribuye T:

$$X_n \sim N(\theta_1, \sigma^2) \quad \Leftrightarrow \quad \overline{X}_n \sim N(\theta_1, \frac{\sigma^2}{n}) \quad \Leftrightarrow \quad \frac{\sqrt{n}\,\overline{X}_n}{\sigma} \sim N(\frac{\theta_1\sqrt{n}}{\sigma}, 1)$$

$$\Leftrightarrow \frac{\sqrt{n}\,\overline{X}_n}{\sigma} - \frac{\theta_1 \cdot \sqrt{n}}{\sigma} \sim N(0,1)$$

Luego sabemos que la distribucion T-student no centralizada se escribe como:

$$\frac{N(0,1) + \mu}{\sqrt{\frac{\chi^2_n}{n}}} \sim F_{n,\mu}$$

Y por teorema visto en la teorica:

$$\frac{S}{\sigma} \sim \sqrt{\frac{\chi^2_{n-1}}{n-1}}$$

$$T = \frac{\sqrt{n}\,\overline{X}_n}{S} = \frac{\frac{\sqrt{n}\,\overline{X}_n}{\sigma}}{\frac{S}{\sigma}} = \frac{\frac{\sqrt{n}\,\overline{X}_n}{\sigma} - \frac{\theta_1 * \sqrt{n}}{\sigma} + \frac{\theta_1\sqrt{n}}{\sigma}}{\frac{S}{\sigma}} \sim F_{n-1,\frac{\theta_1\sqrt{n}}{\sigma}}$$

Finalmente:

$$\boxed{T \sim F_{n-1,\frac{\theta_1\sqrt{n}}{\sigma}}}$$

Vemos que todavia depende de conocer σ pero podriamos pensarlo como que tenemos un valor historico del mismo para este ejercicio, solo por diversion, ya que concretamente, si conocemos el valor de σ habria que utilizar el test de la normal, pero como dijo el DISCLAIMER, lo pensamos antes de que supieramos que este parametro era conocido.

Entonces:

$$\pi_\phi(\theta_1) = \mathbb{P}_{\theta_1}(T \geq k_\alpha) = 1 - F_{n-1,\frac{\theta_1\sqrt{n}}{\sigma}(k_\alpha)}$$

Una vez demostrado esto, es posible calcular la potencia del test $\pi_{\phi_t(\theta_1)}$. Cuando $\alpha = 0{,}05$, $\pi_{\phi_t(\theta_1)} = 0{,}9447$.

Luego del disclaimer, correctamente, si conocemos el valor de σ debemos usar el siguiente test:

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu > 0.$$

$$T = \frac{\sqrt{n}\,\overline{X}_n}{\sigma} \sim N(0,1)$$

$$\phi_n(x) = \mathbb{1}_{\{T \geq k_\alpha\}}$$

$$k_\alpha = z_{1-\alpha} \text{ donde } \sigma \text{ es conocida y vale 1, } P(T \geq z_{1-\alpha}) = 0{,}05$$

$$\pi_{\phi_n}(\theta) = 1 - \Phi\left(k_\alpha + \sqrt{n}\frac{\theta}{\sigma}\right) \text{ donde } \sigma \text{ es conocida y vale 1}$$

$$\pi_{\phi_n}(1) = 1 - \Phi\left(k_\alpha - \sqrt{n}\right) = 0{,}9656$$

Luego para el test del signo:
Sea $X_1, X_2, \ldots, X_n \sim \text{F}$,

$$H_0 : \text{Med(F)} = 0, \text{ es decir, } P(X_i > 0) = P(X_i \leq 0) = 0{,}5,$$
$$H_1 : \text{Med(F)} = \theta_1 > 0.$$

En nuestro caso, la mediana coincide con la media pues F es simétrica respecto de θ. Entonces, testear las hipótesis del test del signo son equivalentes a testear las siguientes:

$$H_0 : \mu = 0,$$
$$H_1 : \mu = \theta_1 > 0.$$

Sea Y_i una variable indicadora definida como:

$$Y_i = \begin{cases} 1 & \text{si } X_i \geq 0, \\ 0 & \text{si } X_i < 0. \end{cases}$$

Notamos que $Y_i \sim \text{Bernoulli}(p)$, donde $p = P(X_i \geq 0)$. Bajo H_0, se tiene que $p = 0{,}5$. La estadística de prueba es la suma de las variables indicadoras:

$$S = \sum_{i=1}^{n} Y_i.$$

Bajo H_0, S sigue una distribución binomial:

$$S \sim \text{Binomial}(n, 0{,}5).$$

$$\phi_s(x) = \mathbb{1}_{\{S \geq k_\alpha\}}$$

la potencia estimada por boostrap cuando $\alpha = 0{,}05$, $\pi_{\phi_s(\theta_1)} = 0{,}7082$.

En conclusión, para testear H0 : $\theta = 0$ contra H1 : $\theta > 0$, con varianza conocida. Luego, el test normal es el mas potente debido a ques es el que mas informacion tiene. Sigue en segundo lugar, el test t el cual tiene una perfomance parecida, aunque para estimar correctamente el valor exacto de la potencia, se deberia hacer mediante estimaciones de la varianza, ya que inicialmente este test esta pensado para σ desconocido. Entonces nos queda, $\pi_{\phi_s(\theta_1)} \leq \pi_{\phi_w(\theta_1)} \leq \pi_{\phi_t(\theta_1)} \leq \pi_{\phi_n(\theta_1)}$ donde

$\pi_{\phi_t(\theta_1)} - \pi_{\phi_s(\theta_1)} = 0{,}2365$ y $\pi_{\phi_t(\theta_1)} - \pi_{\phi_w(\theta_1)} = 0{,}0176$ y $\pi_{\phi_n(\theta_1)} - \pi_{\phi_t(\theta_1)} = 0{,}03846$

3. Conclusión

El desarrollo de este trabajo práctico permitió concoer el test de rango signado de Wilcoxon, un test no paramétrico. Además, la utilización de bootstrap permitió evaluar la potencia del test frente a hipótesis alternativas y compararlos con otros tests ya conocidos, el test t, el test del signo y el test de la normal, mostrando la ventaja del test de la normal cuando se cumplen los supuestos de normalidad.

Referencias

- Test de Wilcoxon de rangos signados, de Elena J. Martínez
- Wilcoxon signed-rank test, wikipedia
- Practical Nonparametric Statistics, de W.J. Conover
- «Statistical Inference Based on Ranks», de Thomas P. Hettmansperger
- Corrección por continuidad
- Noncentral t-distribution